基于多流 CNN-LSTM 网络的群体情绪识别 *

卿粼波，熊文诗?，周文俊，熊珊珊，吴晓红

(四川大学 电子信息学院，成都 610065)

摘 要：群体情绪识别是人机交互领域的前言课题，针对群体情绪识别准确率的问题，结合卷积神经网络（CNN）与长短期记忆网络（LSTM），提出一种多流 CNN-LSTM 网络模型学习群体情绪的静态和动态特征。以视频序列的原始图像、视觉显著图形和叠加的光流图像分别作为三个通道的输入，利用 CNN 网络对空间特征和局部运动特征进行分析，得到的特征图直接输入 LSTM 网络，进行全局运动特征的学习。最后连接 Softmax 分类器，对三个通道的 Softmax 输出进行加权融合，得到分类结果。实验结果表明，本文模型可有效地识别 4 种典型的群体情绪，且识别率高于已有算法，准确度（ACC）和宏平均精度（MAP）分别最高可达 82.6%、84.1%。

关键词：群体情绪识别；卷积神经网络；长短期记忆网络；多流

中图分类号：TP391 doi: 10.3969/j.issn.1001-3695.2017.07.0673

Crowd emotion recognition based on multi-stream CNN-LSTM networks

Qing Linbo, Xiong Wenshi?, Zhou Wenjun, Xiong Shanshan, Wu Xiaohong

(*College of Electronic Information Engineering Sichuan University*, *Chengdu* 610065, *China*)

Abstract: Crowd emotion recognition is a preface topic in human-computer interaction field. Aimed at the problem of the accuracy of group emotion recognition, combined with the convolutional neural network (CNN) and the long and short memory network (LSTM) , this paper developed a multi-stream CNN-LSTM network model to study the static and dynamic characteristics of group emotion. Using the original images, saliency maps and stacked optical images as the input of three channels, the spatial features and local motion features were analyzed using the CNN. In order to learn the global motion information, the output feature maps of CNN were used as the input of LSTM. Finally, connected to the Softmax classifier, weighted fusion was adopted to the output of the three streams Softmax classifier. The experimental results show that the model can effectively identify 4 typical crowd emotions, and the recognition rate is higher than the existing algorithms. The maximum accuracy (ACC) and macro average accuracy (MAP) are up to 82.6% and 84.1%, respectively.

Key Words: crowd emotion recognition; convolutional neural network; long short term memory network; multi-stream

0 引言

人体情绪识别作为智能化人机交互技术中的一个重要组成部分，有着广阔的应用前景。人体情绪识别主要分为个体情绪识别和群体情绪识别两个方面。目前，大多数研究主要集中在基于表情的个体情绪识别问题上，对群体情绪识别的研究相对匮乏。然而，随着城市人口的迅速增长，研究对象由个体逐渐转变为群体，并且在拥挤的环境中，由于遮挡和分辨率的问题，很难根据个人的表情去推断群体的情绪。因此，基于视频的群体情绪识别显得尤为重要，它不仅可以应用于监控视频的异常检测，还可以应用于智慧城市的规划，以给人们提供更加人性化的服务。如何高效地识别群体情绪是目前急需解决的问题。

目前，针对群体情绪识别已有许多基于传统方法的研究，将视频序列帧间的运行特征输入到分类器进行分类。Urizar 等人[1]提出一种基于分层贝叶斯模型的群体情绪识别算法，通过挖掘行为和情绪之间的关系推断群体情绪的状态。Rabiee 等人[2]结合群体行为训练一套基于情绪的 SVM 分类器，对监控视频进行异常检测。Patwardhan[3]对整个视频序列进行边缘检测，并结合网格线性叠加提取特征，利用 SVM 进行分类。Zhang 等人[4]利用结构化轨迹学习检测群体连贯的运动模式，再将运动模式映射到情感平面，最后利用分类器对特征进行分类。尽管上述方法在群体情绪识别中取得了一定的效果，但是由于真实场景的复杂性，不同环境下人工选择的特征量是有差异的，所以模型参数的泛化性能差。

基金项目：成都市科技惠民资助项目（2015-HM01-00293-SF）；中央高校基本科研业务费资助项目（2015SCU04A11）

作者简介：卿粼波（1982-），男，四川成都人，副教授，主要研究方向为图像处理、视频编码；熊文诗（1993-），女（通信作者），硕士研究生，主要研究方向为视频分析、视频编码（xwen_shi@126.com）；周文俊（1992-），男，硕士研究生，主要研究方向为视频分析；熊珊珊（1992-），女，硕士研究生，主要研究方向为视频分析；吴晓红（1970-），女，博士研究生，主要研究方向为图像处理与模式识别、电子通信与系统.

2006 年，Hinton 等人[5]提出了深度学习理论，利用分层抽象的特征提取方式替代了人工选择特征的方法，从而消除人工选择特征的差异性，实现特征的自动学习。典型的深度学习模型卷积神经网络(convolytional neural nets, CNN)以图像的像素值作为输入，通过模拟人脑进行分析学习，在图像识别领域不仅具有较强的鲁棒性且取得了惊人的识别率。针对视频内容的识别，Simonyan 等人[6]认为不仅需要考虑视频静态特征，还应考虑视频的运动特征，因此结合原始静态图像和光流图像提出了双流的的卷积神经网络；除此之外，Zhao 等人[7]利用骨架和原始图两通道信息进行个体的行为识别。易超人等人[8]建立 4 个多层卷及神经网络，学习 4 个不同方向梯度图像的特征，最后融合得到分类结果。长短期记忆网络 (long short term memory network, LSTM) [9]是一种新型的递归神经网络，LSTM 可以对当前的输入有选择的记忆，输出反馈至下一次输入，是一种动态的时间延迟网络，在处理时序相关的的输入时，有着很大的优势。 Donahue 等人[10]针对视频的识别和描述，将 CNN 与 LSTM 结合，提出了长效递归卷积神经网络(long-term recurrent convolutional networks, LRCN)，并取得了较好的识别率； Cai 等人[11]结合 CNN 与双向 RNN 进行面部的表情识别；秦阳等人[12]结合三维卷积神经网络和 LSTM 网络，提出一种融合模型进行行为识别。

深度学习在视频分析中有较高的识别率和泛化能力，但目前还未见有学者将其运用到群体情绪识别中。因此，本文提出一种基于多流 CNN-LSTM 的群体情绪识别网络模型。首先，将视频的原始图像序列、显著图序列和光流序列作为三个通道分别输入 CNN 网络进行特征学习，得到视觉特征；然后，将三个通道的视觉特征分别作为LSTM 网络的输入对全局运动信息进行建模，并利用 Softmax 层进行分类；最后，对三通道 Softmax 层的输出进行平均融合。

1 基于多流 CNN-LSTM 的群体情绪识别网络

1.1 面向群体情绪的视频特征选取

静态图像的识别只需要对图像中的像素值之间的特征进行学习，而针对基于视频的群体情绪识别则需要对视频序列帧间的运行特征进行学习。所以，不仅要考虑空间域上像素值之间的相关性，还要考虑时间域上帧间的相关性。视频承载的多态信息可自然地分解为时空两部分。空间部分由视频的纹理特征组成，对场景和对象的外观进行描述；时间部分体现在帧间的运动特征，描述视频中对象的运动。

如图 1 所示，本文利用视频的原始图像、视觉显著图像和光流图像对视频群体情绪的时空特征进行学习。空间部分，场景及群体的外观特征利用原始视频图像描述。近年来视觉显著图被广泛应用于图像数据处理中，它能够突显出人类视觉系统关注的重点区域，在人群视频图像中，它恰好能体现群体在场景中的显著程度，如图 1(b)所示。本文采用文献[13]中典型的基于多通道的显著图融合方法生产视觉显著图。从时间相关性来看，群体情绪还与群体的运动信息有关，运动平缓表示群体处于放松的状态，运动较激烈则表示群体处于激动的状态或者有异常发生。因此，本文在空间部分，采用表示相邻帧间运动的光流图像[14]对群体的运动特征进行描述。


(a) 原始图像


(b) 视觉显著图像


(c) 光流图像

图 1 面向群体情绪的视频特征

1.2 多流 CNN-LSTM 网络模型

为充分利用 1.1 节所述的视频特征进行群体情绪识别，本文提出如图 2 所示的的多流 CNN-LSTM 网络模型，它由三通道的 CNN-LSTM 网络构成。空间流的两通道分别以原始图像和视觉显著图像作为输入，主要用于静态图像信息的学习。原始图像描述场景和群体静态外观信息，视觉显著图表示群体的显著程度。由于光流图像是根据相邻帧之间的相关性生成的，所以采用光流图像表示视频的局部运动信息，作为时间流的输入。文献[6]提出在一个较短的时间窗口内对光流图像进行连续的叠加，可以更紧凑地表示视频的运行信息，识别效果比原始光流图像好。因此，本文采用叠加光流图像作为输入，叠加的帧数为 10。

CNN 网络不仅有很强的图像特征学习能力，并且在训练时能减少计算量，因此本文采用 3 个 CNN 网络分别对空间流的静态图像信息以及运动流的局部运动信息进行建模。LSTM 网络与 CNN 网络主要的不同之处在于它能够持续保留信息，能够根据之前状态推出后面的状态，从而学习到视频的全局运动特征。为了对视频序列的静态特征、局部运动特征、全局运动特征进行建模，本文以文献[10]为基础融合 CNN 网络与 LSTM 网络，在 CNN 网络第一个全连接层后连接 LSTM 网络。最后对三通道 Softmax 层的输出进行平均融合，得到最终的分类结果。

1.3 网络的选取与训练

本文为了提高群体情绪特征的准确性，采用了 VGG-19[15] 和 AlexNet[16]两个 CNN 网络模型。空间流的两个通道利用 VGG-19 网络模型进行特征的自动学习。首先，采用 ImageNet 数据集[13]对 VGG-19 模型进行预训练；然后，利用用于训练的数据对预训练的模型进行微调。在微调阶段，输入图像的大小统一为 224×224，learing rate 设置为 10^{-4}，momentum 设置为 0.9。为了避免过拟合，本文在每个全连接层后加入 Dropout 层，Dropout_ratio 为 0.5。

文献[18]分别采用 VGG-19 和 Alexnet 模型对叠加光流图像进行训练。实验发现 Alexnet 模型对叠加光流图像的学习能力更强。因此本文时间流的 CNN 采用 AlexNet 模型，根据文献[10]给出的预训练光流模型，利用用于训练的叠加光流序列进

行微调。与空间流 VGG-19 网络参数设置不同的是，运动流 AlexNet 网络的输入图像的大小固定为 227×227，且全连接层后的 Dropout 层的 Dropout_ratio 设置为 0.7，以避免过拟合。

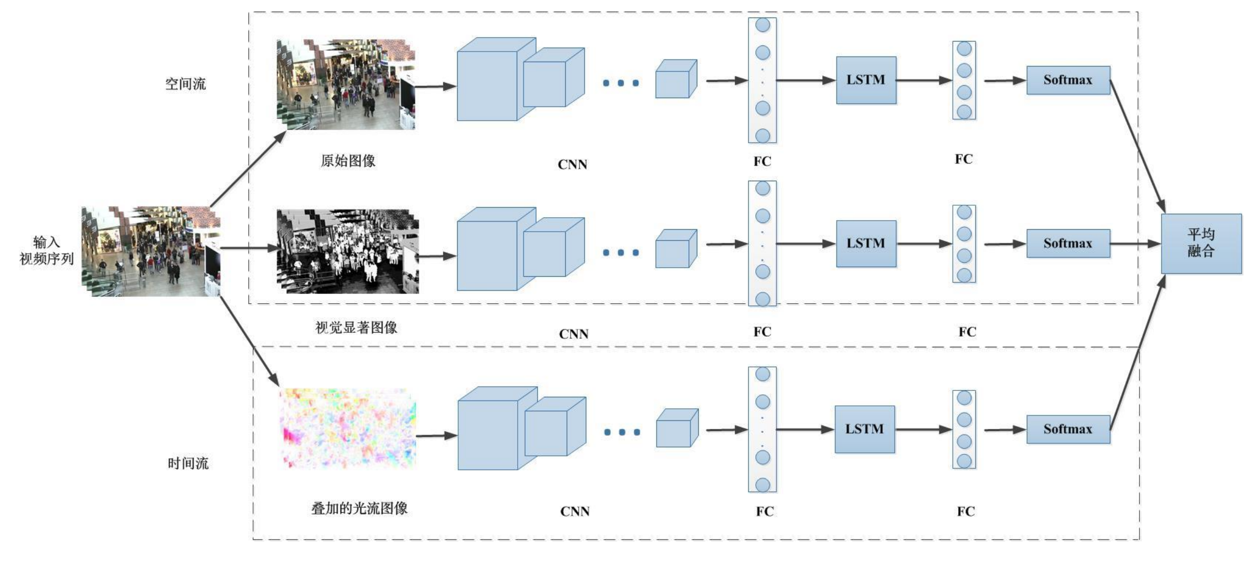


图 2　基于多流 CNN-LSTM 的群体情绪识别网络

CNN 与 LSTM 网络的连接如图 3 所示。LSTM 网络的输入视频帧数 T 被固定为 16 帧，并且每个 LSTM 网络都含有 512 记忆单元，Learing rate 为 10^{-4}，momentum 为 0.9。训练过程如图 4 所示，从图中可以看出，CNN-LSTM 网络在训练开始后很快就成功地从样本数据中学习到了表示群体情绪的时空特征，训练损失函数(train loss)能够得到很好的收敛，在训练迭代 30 000 次后，测试准确度(Test accuary)达到平缓状态，为 80%左右。

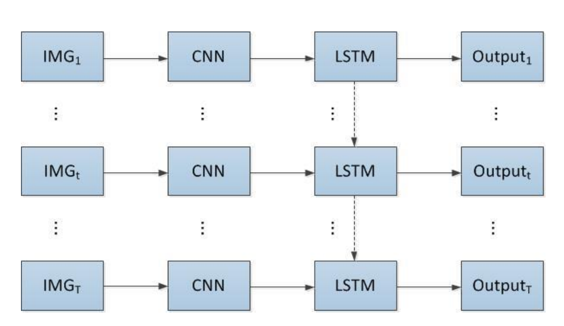


图 3　CNN 网络与 LSTM 网络的连接结构

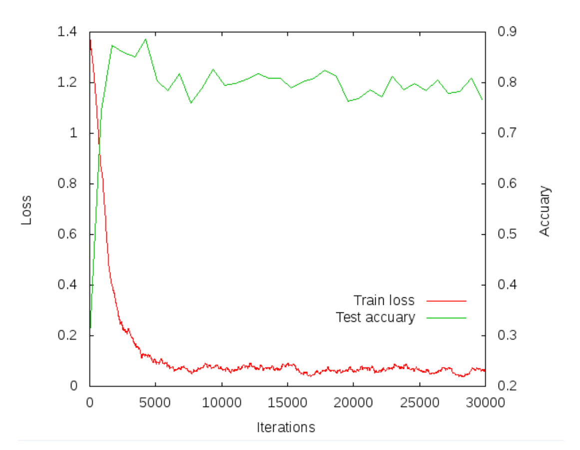


图 4　CNN-LSTM 网络的训练过程

2　实验测试

2.1　实验方法

本文在基于 Python 的深度学习框架 Caffe 环境下进行实验。实验环境如下：

① Intel i5 2.4 GHz 2 Cores；

② NVIDIA GeForce GTX 1070；

③ 8 GB 内存；

④ Ubuntu 14.04×64。

为了评估本文所提的多流 CNN-LSTM 群体情绪识别网络的性能，本文对三通道网络进行了如下实验：

① 只选取原始图像通道。

② 只选取视觉显著图像通道。

③ 只选取叠加光流图像通道。

④ 对原始图像通道和视觉显著图像通道 Softmax 层输出进行平均融合。

⑤ 对原始图像通道和叠加光流图像通道 Softmax 层输出进行平均融合。

⑥ 对原始图像通道、视觉显著图像通道和叠加光流图像通道 Softmax 层输出进行平均融合。

2.2　数据集与评价标准

由于目前关于群体的数据集主要是针对群体行为分析，并没有群体情绪标签的标准数据集，所以本文结合 CUHK 群体数据集[19]、 UCF 数据集[20]、Web 数据集[21]、PET2009 数据集[22]建立具有群体情绪标签的数据集。群体情绪分为 Bored、Excited、Frantic、Relaxed4 个类别。典型的视频场景如图 5 所示。本文采用旋转、加噪声等方法对数据集进行扩展，训练集包含 863 个视频，验证集包含 142 个视频，测试集为文献[4]所采用的测试集，包含 86 个视频。其中，验证集用于训练阶段的测试，当

训练数据集迭代完一次后，则对验证集进行测试，以防止过拟合，测试集则用于对训练好的模型进行测试，验证模型的准确率。

本文采用准确度(accuracy, ACC)和宏平均精确度(macro average precision, MAP)作为评价标准，计算方法如下所示：

$$ACC = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s}(TP_i + FN_i)} \quad (1)$$

$$MAP = \frac{1}{s}\sum_{i=1}^{s} P_i \quad (2)$$

$$P_i = \frac{TP_i}{TP_i + FN_i} \quad (3)$$

其中：s 表示人群情绪类别的数目；P_i 表示第 i 类的精确度；TP_i、FN_i 分别表示第 i 类中正确预测的数目和错误预测的数目。

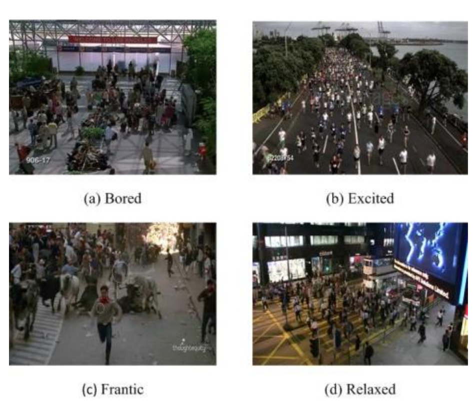


图 5　数据集标签与其对应的典型场景

2.3　实验结果与分析

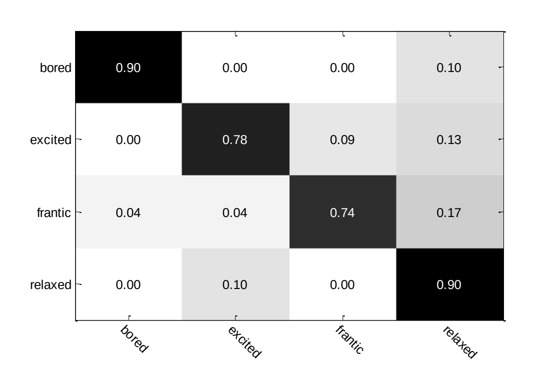
为了说明本文模型的有效性，与文献[4]进行了对比实验。文献[4]采用传统算法对群体情绪进行识别，利用分类器对结构化轨迹学习到的运动特征进行分类。表 1 给出了本文多流 CNN-LSTM 群体情绪识别网络模型与文献[4]群体情绪识别算法的实验对比结果。

从表 1 可以看出，分别以原始图像、显著图形、叠加光流图像训练模型时，原始图像通道的识别结果最好，ACC 和 MAP 分别为 80.2%和 82%，且高于文献[4]的结果。当在原始图像通道中加入显著图像通道进行融合后，由于原始图和显著图都只适用于运动平缓的视频，不能提高运动剧烈的视频的识别率，所以 ACC 和 MAP 有所下降。但在原始图像通道加入叠加光流图像通道进行融合后，ACC 和 MAP 却提高了，这主要是因为原始图像通道对运动平缓的视频（如 Relaxed 类）识别效果好，叠加光流图像通道对运动剧烈的视频（如 Frantic 类）识别效果好，加权融合增强了模型对不同运动程度视频的学习能力。最终，原始图像通道、显著图像通道与叠加光流图像通道融合后，ACC 和 MAP 达到最高，分别为 82.6%、84.1%，与文献[4]相比，分别提高了 7.7%、9.8%。证明了与文献[4]传统算法相比，本文模型具有更准备的分类结果，同时也说明了深度学习在群体情绪上的学习能力。

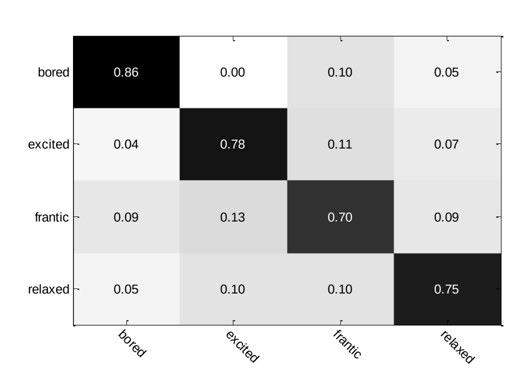
图 6 给出了本文多流 CNN-LSTM 网络模型和文献[4]群体情绪识别结果的混淆矩阵，本文模型选取原始图像通道、显著图像通道与叠加光流图像通道的平均融合结果。从图 6 可以看出，本文模型在 Bored、Frantic、Relaxed 这三类的识别率都高于文献[4]，Excited 与文献[4]相同。从图 4 中的数据集典型场景可以看出 Bored 和 Relaxed 这两类的区分度并不是很大，但是本文模型对 Bored 和 Relaxed 的识别率却最高，因此也证明了本文模型的有效性和准确性，与文献[4]算法相比具有更高的识别率和泛化能力。

表 1　群体情绪识别结果的 ACC 与 MAP

方法		ACC(%)	MAP(%)
文献[4]算法		76.7	76.6
本文模型	原始图	80.2	82.0
	显著图	69.8	78.0
	叠加光流图	38.4	53.1
	原始图+显著图	75.6	78.7
	原始图+叠加光流	81.4	83.0
	三通道融合	**82.6**	**84.1**



a) 本文多流 CNN-LSTM 网络模型



b) 文献[4]算法

图 6　群体情绪识别结果的混淆矩阵

3　结束语

针对机器视觉中群体情绪识别的问题，本文提出了基于多

流 CNN-LSTM 的群体情绪识别网络模型，以视频的原始图像序列、显著图序列和叠加光流序列分别作为三个通道 CNN-LSTM 网络的输入，学习视频中场景和群体的静态特征、群体的局部运动特征和全局运动特征。与已有算法相比，本文多流 CNN-LSTM 网络模型能得到更高的群体情绪识别率，ACC 和 MAP 分别最高可达 82.6%、84.1%。且整个模型基于深度网络，无须先验信息，具有良好的泛化性能。

参考文献：

[1] Urizar O J, Baig M S, Barakova E I, et al. A hierarchical bayesian model for crowd emotions [J]. Frontiers in Computational Neuroscience, 2016, 10 (63): 1-9.

[2] Rabiee H, Haddadnia J, Mousavi H, et al. Emotion-based crowd representation for abnormality detection [J]. arXiv preprint arXiv: 1607. 07646, 2016.

[3] Patwardhan A. Edge based grid super-imposition for crowd emotion recognition [J]. International Research Journal of Engineering and Technology, 2016, 5.

[4] Zhang Y, Qin L, Ji R, et al. Exploring coherent motion patterns via structured trajectory learning for crowd mood modeling [J]. IEEE Trans on Circuits & Systems for Video Technology, 2017, 27 (3): 635-648.

[5] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks [J]. Science, 2006, 313 (5786): 504.

[6] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C]// Advances in Neural Information Processing Systems. 2014: 568-576.

[7] Zhao R Ali H, Smagt PVD. Two-stream RNN//CNN for action recognition in 3D videos [C]// Proc of IEEE//RSJ International Conference on Intelligent Robots and Systems. 2017.

[8] 易超人，邓燕妮. 多通道卷积神经网络图像识别方法 [J]. 河南科技大学学报：自然科学版, 2017, 38 (3): 41-44.

[9] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 2012, 9 (8): 1735-1780.

[10] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2016, 39 (4): 677.

[11] Cai Y, Zheng W, Zhang T, et al. Video based emotion recognition using CNN and BRNN [C]// Proc of Chinese Conference on Pattern Recognition. 2016: 679-691.

[12] 秦阳，莫凌飞，郭文科，等. 3D CNNs 与 LSTMs 在行为识别中的组合及其应用 [J]. 测控技术, 2017, 36 (2): 28-32.

[13] Borji A, Cheng M M, Jiang H, et al. Salient object detection: a benchmark [J]. IEEE Trans on Image Processing, 2015, 24 (12): 5706-5722.

[14] Brox T, Bruhn A, Papenberg N, et al. High accuracy optical flow estimation based on a theory for warping [C]// Proc of European Conference on Computer Vision. 2004: 25-36.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [J]. Computer Science, 2014: 1-14.

[16] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S. l.] : Curran Associates Inc. 2012: 1097-1105.

[17] Berg A, Deng J, Li Feifei. Large scale visual recognition challenge [EB/OL]. (2010) . http://www. image-net. org/challenges/LSVRC/2010/.

[18] Wu Z, Jiang Y G, Wang X, et al. Multi-stream multi-class fusion of deep networks for video classification [C]// Proc of ACM on Multimedia Conference. 2016: 791-800.

[19] Shao J, Chen C L, Wang X. Learning scene-independent group descriptors for crowd understanding [J]. IEEE Trans on Circuits & Systems for Video Technology, 2017, 27 (6): 1290-1303.

[20] Ali S, Shah M. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S. l.] : IEEE Computer Society, 2007: 1-6.

[21] Mehran R, Oyama A, Shah M. Abnormal crowd behavior detection using social force model [C]// Proc of Computer Vision and Pattern Recognition. 2009: 935-942.

[22] Ferryman J, Shahrokni A. PETS2009: dataset and challenge [C]// Proc of the 20th IEEE International Workshop on PERFORMANCE Evaluation of Tracking and Surveillance. 2009: 1-6.